Technical Report OSU-CISRC-10/09-TR50
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210-1277

Ftpsite: **ftp.cse.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2009**
File: **TR50.pdf**
Website: **http://www.cse.ohio-state.edu/research/techReport.shtml**

Sequential Organization of Speech in Reverberant Environments by Integrating Monaural Grouping and Binaural Localization

John Woodruff
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210, USA
woodrufj@cse.ohio-state.edu

DeLiang Wang
Department of Computer Science and Engineering & Center for Cognitive Science
The Ohio State University, Columbus, OH 43210, USA
dwang@cse.ohio-state.edu

Abstract – Existing binaural approaches to speech segregation place an exclusive burden on location information. These approaches can achieve excellent performance in anechoic conditions but degrade rapidly in realistic environments where room reverberation corrupts localization cues. In this work we propose to integrate monaural and binaural processing to achieve sequential organization and localization of speech in reverberant environments. The proposed approach builds on monaural analysis for simultaneous organization, and combines it with a novel method for generation of location-based cues in a probabilistic framework that jointly achieves localization and sequential organization. We compare sequential organization performance against a model-based system that uses only monaural cues and an exclusively binaural system, and localization performance against two existing methods that do not utilize monaural grouping. Results suggest that the proposed integration of monaural grouping and binaural localization allows for improved source localization and robust sequential organization performance in environments with considerable reverberation.

Index Terms – Sequential organization, binaural sound localization, monaural grouping, room reverberation.

1 Introduction

Most existing approaches to binaural or sensor-array based speech segregation have relied exclusively on directional cues embedded in the differences between signals recorded by multiple microphones [3, 31]. These approaches may be characterized as spatial filtering (or beamforming), which enhances the signal from a specific direction. Spatial filtering approaches can be very effective in certain acoustic conditions. On the other hand, beamforming has well known limitations. Chief among them is substantial performance degradation in reverberant environments. Rigid surfaces reflect a sound source incident upon them, hence corrupting directional cues [5].

In spite of the degradation of localization cues in reverberant environments, human listeners are able to effectively localize multiple sound sources in such environments [16] and use localization cues as one mechanism in *auditory scene analysis* (ASA) [4]. This perceptual ability continues to motivate binaural approaches that attempt to utilize localization cues in a manner that is robust to room reverberation. In this work we propose a framework that integrates monaural and binaural analysis to achieve robust localization and segregation of speech in reverberant environments.

In the language of ASA, the segregation problem is one of grouping sound components of the mixture across frequency and across time into *streams* associated with the individual sources. The terms simultaneous organization and sequential organization are used to refer to grouping across frequency and across time, respectively. Our proposed system achieves simultaneous organization using monaural cues. This allows locally extracted, unreliable binaural cues to be integrated across frequency and short, continuous time intervals. This integration enhances the robustness of localization cues in reverberant conditions, and robust localization cues are in turn used to achieve sequential organization. Our computational framework is partly motivated by psychoacoustic studies suggesting that binaural cues may not play a dominant role in simultaneous organization, but are important for sequential organization [10–12]. Our approach marks a significant departure from the dominant paradigm in binaural segregation that relies solely on directional cues to achieve both simultaneous and sequential organization.

Prior work exploring the integration of monaural and binaural cues is limited. In [25], localization cues are used to perform initial segregation in reverberant conditions. Initial segregation provides a favorable starting point for estimating the pitch track of the target voice, which is then used to further enhance the target signal. In [34], pitch and interaural time difference (ITD) are used jointly in a recurrent timing neural network to achieve speech segregation, but the focus is on speech in anechoic environments. In [9], pitch and ITD are used to achieve localization of simultaneous speakers in reverberant environments. Our prior work analyzes the impact of idealized monaural grouping on localization and segregation of speech in reverberant environments [33]. The algorithm proposed here takes advantage of recent developments in monaural segregation of voiced speech [17], provides a novel method

for the generation of location-dependent binaural cues and integrates binaural cues within a probabilistic framework to achieve localization and sequential organization.

Sequential organization is a challenging problem in speech segregation. Grouping components of a mixture across disparate regions of time to form a cohesive stream associated with a source is vital for real-world deployment of a segregation system. Using binaural cues is attractive because monaural features alone may not be able to solve the problem. For example, in a mixture of two male speakers who have a similar vocal range, pitch-based features cannot be used to group components of the mixture that are far apart in time. As a result, feature-based monaural systems have largely avoided sequential organization by focusing on short utterances of voiced speech [29] or assuming prior knowledge of the target signal's pitch [20], or achieved sequential organization by assuming speech mixed with non-speech interference [19].

Shao and Wang explicitly addressed sequential organization in a monaural system using a model-based approach [27]. In this work they use feature-based monaural processing to perform simultaneous organization of voiced speech, and a speaker recognition-based approach to perform sequential organization of the already formed time-frequency segments. They provide extensive results on sequential organization performance in co-channel speech mixtures as well as speech mixed with non-speech intrusions. However, they do not address sequential organization in reverberant environments, and mismatch between training and testing conditions is known to cause performance degradation in model-based systems [7].

In the following section we provide an overview of the proposed architecture. In Section 3 we discuss monaural simultaneous organization of voiced speech. Section 4 outlines our methods for extraction of binaural cues and for calculating azimuth-dependent cues, and a mechanism for weighting cues based on their expected reliability. In Section 5, we formulate joint sequential organization and localization in a probabilistic framework. We assess both localization and sequential organization performance, and compare the proposed system to existing methods in Section 6. We conclude with a discussion in Section 7.

2 System Overview

The proposed system uses both monaural and binaural processing to achieve sequential organization of voiced speech. A diagram is provided in Figure 1. The input to the system is a binaural recording of a speech source mixed with one or more interfering signals. The recordings are assumed to be made using a dummy head with two microphones inserted in the ears, such as KEMAR [14]. Since a human listener is implied, we will refer to the two mixture signals as the left ear and right ear signals, denoted by $l[n]$ and $r[n]$ respectively.

Both signals are first passed through a bank of 128 gammatone filters [23] with center frequencies from 50 to 8000 Hz spaced on the equivalent rectangular bandwidth scale, where we denote the signals for frequency channel c as $l_c[n]$ and $r_c[n]$. Each filtered signal is processed using 20-ms time frames with a frame shift of 10-ms to create a *cochleagram* of time-frequency

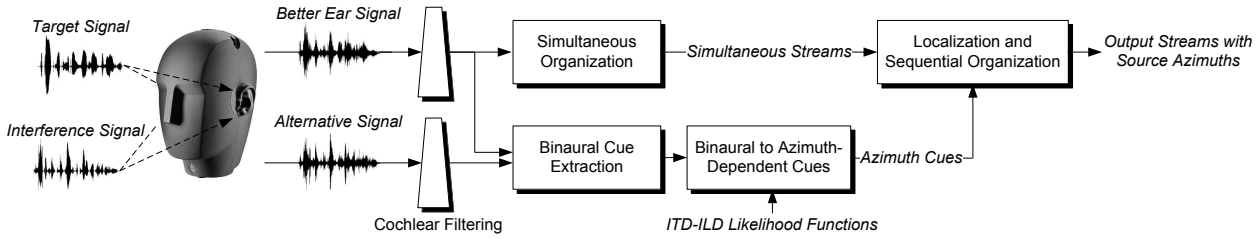


Figure 1: Diagram of the proposed system. Source signals are assumed to be recorded by a binaural microphone and are fed as input to the system. Cochlear filtering is applied to both signals. Monaural processing generates simultaneous streams from the *Better Ear Signal*. Both signals are used to generate azimuth cues. Simultaneous streams and azimuth cues are combined in the final localization and sequential organization stage.

(T-F) units [31].

In the first stage of the system, the tandem algorithm of Hu and Wang [17, 18] is used to process the *better ear* signal and form simultaneous streams. By better ear signal, we mean the signal in which the input SNR is higher. A simultaneous stream refers to a collection of T-F units over a continuous time interval that are thought to be dominated by the same source signal. The tandem algorithm performs simultaneous organization of voiced speech using monaural cues such as harmonicity and amplitude modulation. Unvoiced speech presents a greater challenge for monaural systems and is not dealt with in this study (see [19]).

Binaural cues are extracted that measure differences in timing and level between corresponding T-F units of the left and right ear signals. A set of trained, azimuth-dependent likelihood functions are then used to map from timing and level differences to cues related to source location. Azimuth cues are integrated over simultaneous streams in a probabilistic framework to achieve sequential organization and to estimate the underlying source locations. The output of the system is a set of streams, one for each source in the mixture, and the azimuth angles of the underlying sources.

3 Simultaneous Organization

Simultaneous organization in computational auditory scene analysis (CASA) systems forms simultaneous streams, each of which may contain disconnected T-F segments across a continuous time interval. We use the tandem algorithm proposed in [17, 18] to generate simultaneous streams for voiced regions of the better ear mixture. The tandem algorithm iteratively estimates a set of pitch contours and associated simultaneous streams. In a first pass, T-F segments that contain voiced speech are identified using cross-channel correlation of correlogram responses. Up to two pitch points per time frame are estimated by finding peaks in the summary correlogram created from only the selected, voiced T-F segments. For each pitch point found, T-F units that are consistent with that pitch are identified using a set of trained multi-layer perceptrons (one for each frequency channel). Pitch points and associated

sets of T-F units are linked across time to form pitch contours and associated simultaneous streams using a continuity criterion that measures pitch deviation and frequency overlap. Pitch contours and simultaneous streams are then iteratively refined until convergence.

We focus on two talker mixtures in reverberant environments, and find that in this case the continuity criterion used in the tandem algorithm for connecting pitch points and simultaneous streams across time is too liberal. We find that performance improves if we break pitch contours and simultaneous streams when the pitch deviation between time frames is large. Specifically, let τ_1 and τ_2 be pitch periods from the same contour in neighboring time frames. If $|\log_2(\tau_1/\tau_2)| > 0.08$, the contour and associated simultaneous streams are broken into two contours and two simultaneous streams. The value of 0.08 was selected on the basis of informal analysis, and was not specifically tuned for optimal performance on the data set discussed in Section 6.

An example set of pitch contours and simultaneous streams are shown in Figure 2 for a mixture of two talkers in a reverberant environment with 0.4 sec. reverberation time (T_{60}). There are a total of 27 contour and simultaneous stream pairs shown. The cochleagram of the mixture is shown in 2(a). In 2(b), detected pitch contours are shown by alternating between circles and squares, while ground truth pitch points generated from the pre-mixed reverberant signals are shown as solid lines. In Figure 2(c), each gray level corresponds to a separate simultaneous stream. One can see that simultaneous streams may contain multiple segments across frequency but are continuous in time.

4 Binaural Processing

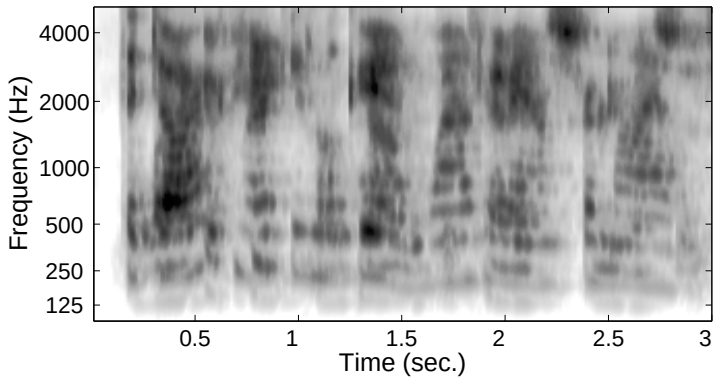
In this section we describe how binaural cues are extracted from the mixture signals and propose a mechanism to translate these cues into information about the azimuth of the underlying source signals. We also discuss a method to weight binaural cues according to their expected reliability.

4.1 Binaural Cue Extraction

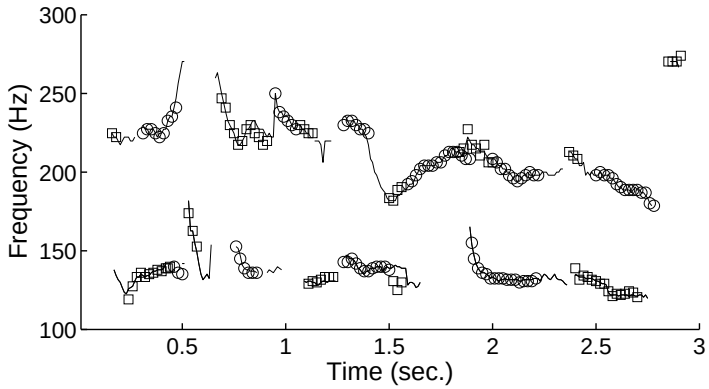
Two primary binaural cues used by humans for localization of sound sources are interaural time difference (ITD) and interaural level difference (ILD) [2]. We calculate ITD in individual frequency bands using the normalized cross-correlation by first computing,

$$C(c,m,\tau) = \frac{\sum_{n=0}^{T_n-1} l_c[m\frac{T_n}{2}-n]r_c[m\frac{T_n}{2}-n-\tau]}{\sqrt{\sum_{n=0}^{T_n-1}(l_c[m\frac{T_n}{2}-n])^2}\sqrt{\sum_{n=0}^{T_n-1}(r_c[m\frac{T_n}{2}-n-\tau])^2}},$$

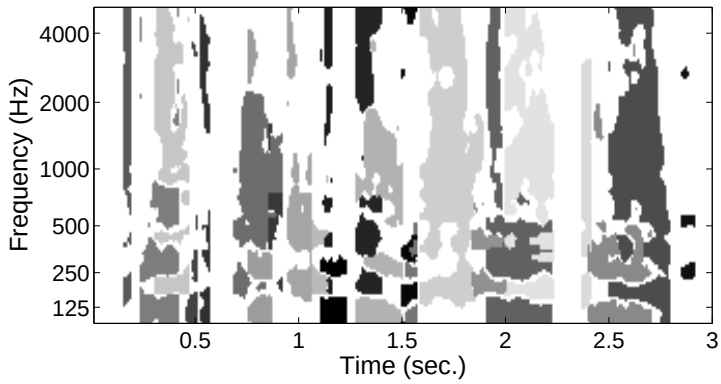
where τ is the time lag for the correlation, c and m index frequency channels and time frames, respectively, and T_n denotes the number of samples per time frame. The ITD is then defined as the time lag that produces the maximum peak in the normalized cross-correlation function,



(a) Cochleagram



(b) Detected pitch contours



(c) Simultaneous streams

Figure 2: Example of multi-pitch detection and simultaneous organization using the tandem algorithm. (a) Cochleagram of a two-talker mixture. (b) Ground truth pitch points (solid lines) and detected pitches (circles and squares). Different pitch contours are shown by alternating between circles and squares. (b) Simultaneous streams corresponding to different pitch contours are shown with varying gray-scale values.

or,

$$\tau_{c,m} = \arg\max_{\tau \in \mathrm{T}} C(c, m, \tau), \tag{1}$$

where T denotes the set of peaks in $C(c, m, \tau)$.

ILD corresponds to the energy ratio in dB between the two signals in corresponding T-F units.

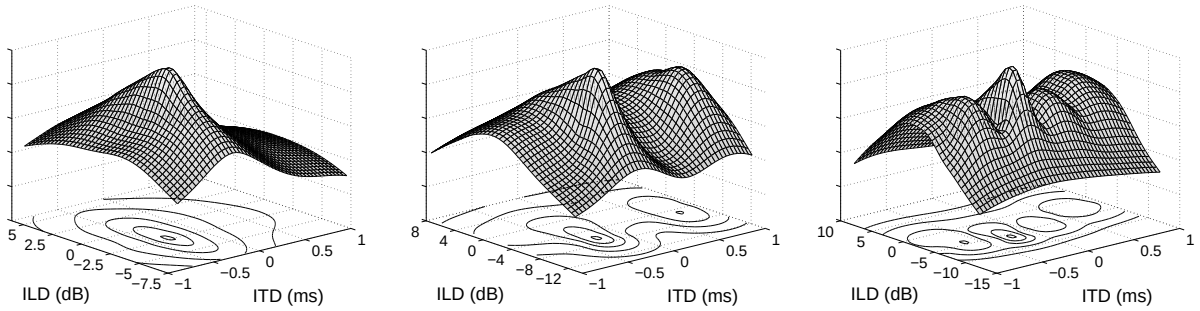
$$\lambda_{c,m} = 10 \log_{10} \left(\frac{\sum_{n=0}^{T_n - 1} (l_c[m\frac{T_n}{2} - n])^2}{\sum_{n=0}^{T_n - 1} (r_c[m\frac{T_n}{2} - n])^2} \right). \tag{2}$$

4.2 Azimuth-Dependent ITD-ILD Likelihood Functions

If one assumes binaural sensors in an anechoic environment, a given source position relative to the listener's ears will produce a specific, frequency dependent set of ITDs and ILDs for that listener. In order to effectively integrate information across frequency for a given source position, these patterns must be taken into account. Further, integration of ITD and ILD cues extracted from reverberant mixtures of multiple sources should account for deviations from the free-field patterns.

In this work we focus on a subset of possible source locations. Specifically, we restrict the source signals to be in front with elevation of 0°. As a result, source localization reduces to azimuth estimation in the interval $[-90°, 90°]$. To translate from raw ITD-ILD information to azimuth, we train a joint ITD-ILD likelihood function, $P_c(\tau_{c,m}, \lambda_{c,m}|\phi)$, for each azimuth, ϕ, and frequency channel, c. Likelihood functions are trained on single-source speech in various room configurations and reverberation conditions using kernel density estimation [28]. The room size, listener position, source distance to listener and reflection coefficients of the wall surfaces are randomly selected from a pre-defined set of 540 possibilities. An ITD-ILD likelihood function is generated for each of 37 azimuths, $[-90°, 90°]$ spaced by 5°, and 128 frequency channels with center frequencies from 50 to 8000 Hz. With these functions, we can translate the ITD-ILD cues in a given T-F unit into an azimuth-dependent likelihood curve. Due to reverberation, we do not expect the maximum of the likelihood curve in each T-F unit to be a good indication of the dominant source's azimuth, but hope that a good indication of the dominant source's azimuth emerges through integration over a simultaneous stream.

The likelihood distributions capture the frequency dependent pattern of ITDs and ILDs for a specific azimuth and the multi-peak ambiguities present at higher frequencies. Each distribution has a peak corresponding to the free-field cues for that angle, but also captures common deviations from the free-field cues due to reverberation. We show three distributions in Figure 3 for azimuth 25°. Note that, in addition to the above points, the azimuth-dependent distributions capture the complementary nature of localization cues [2] in that ITD provides greater discrimination between angles at lower frequencies (note the large ILD variation in the 400 Hz example) and ILD provides greater discrimination between angles at higher frequencies (note the large ITD variation in the 2500 Hz example).



(400 Hz) (1000 Hz) (2500 Hz)

Figure 3: Examples of ITD-ILD likelihood functions for azimuth 25° at frequencies of 400, 1000 and 2500 Hz. Each example shows the log-likelihood as a surface with projected contour plots that show cross sections of the function at equally spaced intervals.

Our approach is adapted from the one proposed in [24]. In that system ITD and ILD are directly used to indicate whether or not the target source is dominant in each T-F unit. The method is developed with a simple idea in mind: an ITD-ILD pair measured from the mixture will be close to the target's free-field ITD-ILD when the target is dominant. To utilize this observation, Roman et al. train two ITD-ILD likelihood functions for each frequency channel, $P_c(\tau_{c,m}, \lambda_{c,m}|H_0)$ and $P_c(\tau_{c,m}, \lambda_{c,m}|H_1)$, where H_0 denotes the hypothesis that the target signal is stronger than the interference signal in unit $u_{c,m}$, and H_1 that the target is weaker. Mixtures are used for training where the pre-mixed target and interference signals are used to determine if a T-F unit satisfies H_0 or H_1. The distributions $P_c(\tau_{c,m}, \lambda_{c,m}|H_0)$ and $P_c(\tau_{c,m}, \lambda_{c,m}|H_1)$ are trained for each target/interference angle configuration. The ITD search space is limited around the expected free-field target ITD in both training and testing to avoid the multi-peak ambiguity in higher frequency channels. For a test utterance, the azimuths of both target and interference sources are estimated. Given these angles, the appropriate set of likelihood distributions is selected. ITDs (within the limited search range) and ILDs are calculated for each T-F unit of the mixture and the maximum a posteriori decision rule is used to estimate a binary mask for the target source.

There are two primary reasons for altering the method in [24] to the one proposed here. First, our proposed approach lowers the training burden because likelihood functions are trained for each angle individually, rather than as combinations of angles. Second, the fact that we do not limit the ITD search space in training allows us to use the likelihood functions in estimation of the underlying source azimuths, rather than requiring a preliminary stage to estimate the angles. We show in Section 6.2 that our proposed localization method, which utilizes the ITD-ILD likelihood functions, performs significantly better than the method proposed in [24].

Because we do not limit the ITD search space, our approach does not attempt to resolve the multi-peak ambiguity inherent in high frequency ITD calculation at the T-F unit level. The lack of one-to-one mapping from phase to time in the high frequency channels is captured in the

likelihood functions (see Figure 3). The ambiguity between sources originating from different azimuths is naturally resolved when integrating across frequency within a simultaneous stream.

4.3 Cue Weighting

In reverberant recordings, many T-F units will contain cues that differ significantly from free-field cues. Although these deviations are incorporated in the training of the ITD-ILD likelihood functions described above, including a weighting function or cue selection mechanism that indicates when an azimuth cue should be trusted can improve localization performance. Motivated by the *precedence effect* [21], we incorporate a simple cue weighting mechanism that identifies strong onsets in the mixture signal. When a large increase in energy occurs, and shortly thereafter, the azimuth cues are expected to be more reliable. We therefore generate a weight, $w_{c,m}$, associated with $u_{c,m}$ that measures the change in signal energy over time. First, we define a recursive method to measure the average signal energy in both left and right channels as follows,

$$e_c[n] = \alpha(l_c[n]^2 + r_c[n]^2) + (1-\alpha)e_c[n-1]. \tag{3}$$

Here $\alpha \in [0,1]$ and $\alpha = \frac{1}{Tf_s}$, where T denotes the time constant for integration and f_s is the sampling frequency of the signals. We set $T = 10$-ms in this study. We then calculate the percent of change in energy between samples and average over an integration window to get,

$$w_{c,m} = \frac{1}{T_n}\sum_{n=0}^{T_n-1}\frac{e_c[m\frac{T_n}{2}-n] - e_c[m\frac{T_n}{2}-n-1]}{e_c[m\frac{T_n}{2}-n-1]}. \tag{4}$$

$w_{c,m}$ is then normalized over each mixture to have values between 0 and 1.

We have found measuring change in energy using this method to provide better results than simply taking the change in average energy from unit to unit, or taking the more traditional derivative of the signal envelope [31]. We have also found better performance by keeping only those weights above a specified threshold. The difficulty with a fixed threshold however, is that one may end up with a simultaneous stream with no unit above the threshold. To avoid this we set a threshold for each simultaneous stream so that the T-F units exceeding the threshold retain 25% of the signal energy in the simultaneous stream. We have found that the system is not particularly sensitive to the value of 25% and that values between about 15% and 40% give similar performance in terms of sequential organization.

Alternative selection mechanisms have been proposed in the literature [9,13,32]. Faller and Merimaa proposed *interaural coherence* (IC) as a cue selection mechanism [13]. IC is defined as the maximal value of the cross-correlation function, or, $C(c,m,\tau_{c,m})$. They suggest that when the IC value is high for a given T-F unit, then the binaural cues associated with that unit can be trusted, and propose a thresholding mechanism to select T-F units with reliable cues. In preliminary experiments we found the proposed method to outperform selection

methods based on IC as a high IC value does not necessarily ensure a reliable ITD and ILD. Room acoustics are typically modeled using linear, time-invariant filters. If a sinusoidal signal is passed through such a filter, the output is also sinusoidal but with altered phase and amplitude. As a result, reverberant left and right room impulse response functions for a certain azimuth can produce perfectly coherent sinusoids with a different relative ITD and ILD than anechoic left and right room impulse response functions for the same azimuth. This produces T-F units that have a high IC value with ITD and ILD cues that do not indicate the azimuth of the underlying source.

The method proposed in [32] uses ridge regression to learn a finite-impulse response filter that predicts localization precision for single-source reverberant speech in stationary noise. This method essentially identifies strong signal onsets, as does our approach, but requires training. The study in [9] finds that a precedence motivated cue weighting scheme performs about as well as two alternatives on a database of two talkers in a small office environment.

5 Localization and Sequential Organization

As described above, the first stage of the system generates simultaneous streams for voiced regions of the better ear mixture and extracts azimuth-dependent cues for all T-F units using the left and right ear mixtures. In this section we describe the source localization and sequential organization process. The goal of this stage is to correctly label the simultaneous streams as being target or interference dominant and to estimate the azimuths of the underlying sources. Our approach jointly determines the source angles and simultaneous stream labels in a maximum likelihood framework, which is inspired by the model-based sequential organization scheme proposed in [26].

Let N be the number of sources in the mixture, and I be the number of simultaneous streams formed using monaural analysis. Denote the set of all possible azimuths as Φ and the set of simultaneous streams as $S = \{s_1, s_2, ..., s_I\}$. Let Y be the set of all N^I sequential organizations, or labelings, of the set S and y be a specific organization. We seek to maximize the joint probability of a set of angles and a sequential organization given the observed data, D. This can be expressed as,

$$\hat{\phi}_0, \ldots, \hat{\phi}_{N-1}, \hat{y} = \underset{\phi_0,\ldots,\phi_{N-1}\in\Phi, y\in Y}{\arg\max} P(\phi_0, \ldots, \phi_{N-1}, y|D). \tag{5}$$

For simplicity, assume that $N = 2$ and apply Bayes rule to get,

$$\begin{aligned} \hat{\phi}_0, \hat{\phi}_1, \hat{y} &= \underset{\phi_0,\phi_1\in\Phi, y\in Y}{\arg\max} \frac{P(D|\phi_0, \phi_1, y)P(\phi_0, \phi_1, y)}{P(D)}, \\ &= \underset{\phi_0,\phi_1\in\Phi, y\in Y}{\arg\max} P(D|\phi_0, \phi_1, y), \end{aligned} \tag{6}$$

assuming that all angles and sequential organizations are equally likely.

Now, let S_0 be the set of simultaneous streams associated with ϕ_0 and S_1 be the set of simultaneous streams associated with ϕ_1 by y. Using ITD and ILD as the observed mixture data, and assuming independence between simultaneous streams and between T-F units of the same simultaneous stream, we can express Equation (6) as,

$$\hat{\phi}_0, \hat{\phi}_1, \hat{y} = \underset{\phi_0,\phi_1 \in \Phi, y \in Y}{\arg\max} \Big(\prod_{s_i \in S_0} \prod_{u_{c,m} \in s_i} P_c(\tau_{c,m}, \lambda_{c,m} | \phi_0) \cdot \prod_{s_j \in S_1} \prod_{u_{c,m} \in s_j} P_c(\tau_{c,m}, \lambda_{c,m} | \phi_1) \Big)$$

One can express the above equation as two separate equations that can be solved simultaneously in one polynomial-time operation as,

$$\hat{\phi}_0, \hat{\phi}_1 = \underset{\phi_0,\phi_1 \in \Phi}{\arg\max} \Big(\sum_{i=1}^{I} \max_{k \in \{0,1\}} \Big(\sum_{u_{c,m} \in s_i} \log(P_c(\tau_{c,m}, \lambda_{c,m} | \phi_k)) \Big) \Big), \tag{7}$$

$$\hat{y}_i = \underset{k \in \{0,1\}}{\arg\max} \Big(\sum_{u_{c,m} \in s_i} \log(P_c(\tau_{c,m}, \lambda_{c,m} | \phi_k)) \Big), \tag{8}$$

where $\hat{y}_i$ denotes the label of s_i. The key observation in moving to Equations (7) and (8) is that the majority of sequential organizations in the set Y cannot maximize the likelihood as expressed in Equation (7). For a given set of angles, if one assumes independence between simultaneous streams as in Equation (7), then the sequential organization that maximizes the likelihood is the organization where the likelihood of each simultaneous stream is maximized independently.

Incorporating the weighting parameter defined in Section 4.3, Equations (7) and (8) become,

$$\hat{\phi}_0, \hat{\phi}_1 = \underset{\phi_0,\phi_1 \in \Phi}{\arg\max} \Big(\sum_{i=1}^{I} \max_{k \in \{0,1\}} \Big(\sum_{u_{c,m} \in s_i} w_{c,m} \log(P_c(\tau_{c,m}, \lambda_{c,m} | \phi_k)) \Big) \Big), \tag{9}$$

$$\hat{y}_i = \underset{k \in \{0,1\}}{\arg\max} \Big(\sum_{u_{c,m} \in s_i} w_{c,m} \log(P_c(\tau_{c,m}, \lambda_{c,m} | \phi_k)) \Big). \tag{10}$$

For the case with $N > 2$, use $k \in \{0, 1, \ldots, N-1\}$ rather than $k \in \{0, 1\}$ in both Equations (9) and (10). The complexity of the search space is $I|\Phi|^N$, which is reasonable when the number of sources of interest is relatively small and the size of the azimuth space is moderate. In our experiments in Section 6, $|\Phi| = 37$ and $N = 2$.

6 Evaluation and Comparison

In this section we evaluate source localization and sequential organization of the proposed system. We analyze localization performance with and without the proposed weighting mechanism and compare the proposed method to two existing methods in various reverberation conditions. We also evaluate sequential organization performance in various reverberation conditions and compare to a model-based approach and an exclusively binaural approach.

6.1 ITD-ILD Likelihood Training and Mixture Generation

We use the ROOMSIM package [8] to generate impulse responses that simulate binaural input at human ears. This package uses measured *head-related transfer function* (HRTF) data from a KEMAR dummy head [14] in combination with the image method for simulating room acoustics [1]. We generate a training and a testing library of binaural impulse responses for 37 direct sound azimuths between -90° and 90° spaced by 5°, and 7 T_{60} values between 0 and 0.8 seconds. In the training library, 3 room size configurations, 3 source distances from the listener and 5 listener positions in the room are used. In the testing library, 2 room size configurations (different from those in training), 3 source distances from the listener and 2 listener positions (different from those in training) are used. For training the ITD-ILD likelihood distributions, speech signals randomly selected from the TIMIT database [15] are convolved with a randomly selected impulse response pair from the training library (for a specified angle). Training is performed over 100 reverberated signals for each of the 37 azimuths.

For all testing mixtures we select both a target and interference speech signal from the TIMIT database, pass the signals through an impulse response pair from the testing library for a desired azimuth and room reverberation time, and sum the resulting binaural target and interference signals to create a binaural mixture. Two hundred mixtures are generated for each of 7 reverberation conditions: T_{60} = 0, 0.1, 0.2, 0.3, 0.4, 0.6, 0.8 seconds. Each mixture contains two source signals, where a room, source distance and listener position are randomly selected and applied to both sources, and source azimuths are selected randomly to be between 10° and 120° apart. The average azimuth spacing over each set of 200 mixtures is about 53°. Speech utterances, azimuths and room conditions remain constant across different T_{60} times. Only the reflection coefficient of the wall surfaces was changed to achieve the selected T_{60}. The SNR of each mixture is set to 0 dB using the dry, monaural TIMIT utterances. This results in better ear mixtures that average 2.8 dB in anechoic conditions down to 1 dB in 0.8 sec. T_{60}. Mixture lengths are determined using the target utterance with the interference signal either truncated or concatenated with itself to match the target length. In order to make a comparison to the model-based approach (discussed further in Section 6.3), the speakers used for the test mixtures are drawn from the set of 38 speakers in the DR1 dialect region of the TIMIT training database.

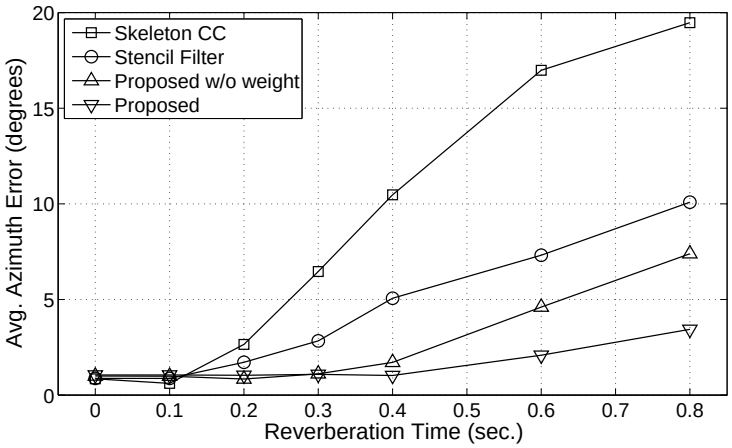


Figure 4: Azimuth estimation error averaged over 200 mixtures, or 400 utterances, for various reverberation times. Results are shown using the proposed approach with and without cue weighting, and two alternative approaches.

6.2 Localization Performance

In this section we analyze the localization accuracy of the method described in Section 5. Specifically, we measure average azimuth estimation error with or without cue weighting. We also compare localization performance to two existing methods for localization of multiple sound sources, as proposed in [22, 24].

The approach proposed by Liu et al. in [22], termed the *stencil filter*, uses Fourier analysis and assumes two omni-directional microphones. ITDs are detected for each frequency bin and time frame and are counted as evidence for a particular azimuth if the ITD falls along the angle's "primary" or "secondary" traces. The primary trace is simply the predicted ITD for that angle, which is a constant function of frequency if one assumes omni-directional microphones. The secondary traces are due to the time/phase ambiguity at higher frequencies where the wavelengths of the signal are shorter than the distance between microphones. For comparison on the database described, some changes were necessary to account for the (somewhat) frequency-dependent nature of ITDs as detected by a binaural system and the discrete azimuth space. Further, because angles are assumed constant over the length of the mixture, azimuth responses from the stencil filter were integrated over all time frames for added accuracy and the two most prominent peaks were selected as the underlying source angles.

The method proposed in [24] computes a "skeleton" cross-correlogram of the mixture signal in which the time-lag dimension is warped to azimuth using a learned set of monotonic functions. The response is then integrated across time and frequency and again, the two most prominent peaks in the response are selected as the underlying source angles. The skeleton aspect of the approach narrows broad peaks in a cross-correlogram in order to increase the estimation resolution, primarily for when more than one source azimuth is being estimated.

Average azimuth error using all four approaches is shown in Figure 4. Estimation is performed for 400 source signals (2 in each of 200 mixtures) and for 7 reverberation times. The results indicate that including weights associated with signal onsets improves azimuth estimation when significant reverberation is present. We can also see that both proposed methods outperform the existing methods for reverberation times of 200 ms or larger. The improvement relative to the skeleton cross-correlogram method is over 15° in highly reverberant conditions, and over 5° relative to the stencil filter approach. In anechoic conditions and 0.1 sec. T_{60}, all methods achieve roughly 1° average error.

The advantage of the proposed system is that azimuth-dependent cues are not integrated over the entire mixture, as they are in the two existing systems used for comparison. The combination of monaural grouping, and localization within the sequential organization framework integrates azimuth-dependent cues over a subset of the mixture in which a single source is considered dominant. In this way, voiced speech segregation and localization are jointly achieved. We found that simply integrating the azimuth cues across time and frequency to form an azimuth-dependent curve and selecting the two most prominent peaks yielded performance comparable to the stencil filter approach.

6.3 Sequential Organization Performance

The primary task of our proposed system is to perform sequential organization, or to generate a set of source-dependent labels for the simultaneous streams formed using monaural analysis. To measure performance on this task we use the *ideal binary mask* (IBM). The IBM has been proposed as the computational goal of CASA systems [30] and has been shown to dramatically improve speech intelligibility when applied to noisy mixtures [6]. The IBM is a binary labeling of mixture T-F units such that when target energy is stronger than interference energy, the T-F unit is labeled with 1, and when target energy is weaker, the T-F unit is labeled with 0. Note that the IBM labels not only T-F units corresponding to voiced speech, but also those corresponding to unvoiced speech. We evaluate sequential organization performance by finding the percentage of mixture energy contained in the simultaneous streams that is correctly labeled by the proposed approach, where ground truth labeling of a T-F unit in a simultaneous stream is generated using the IBM of the better ear mixture. We measure the mixture energy in dB.

We compare performance against two "ceiling" measures that incorporate ideal knowledge, to a recent model-based system [27], and to a system that exclusively uses binaural cues. We refer to the first ceiling performance measure as *ideal sequential organization* (ISO). In this case, a target/interference decision is made for each simultaneous stream based on whether the majority of the mixture energy is labeled target or interference by the IBM. We refer to the second ceiling performance measure as *ground truth pitch* (G.T. Pitch). In this case, we label each simultaneous stream based on the proximity of its associated pitch contour

Table 1: Percent of correctly labeled simultaneous stream energy in various reverberation conditions.

$\mathbf{T_{60}}$ **(sec.)**	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.6**	**0.8**
ISO	88.0	88.4	88.7	88.9	88.7	88.2	87.8
G.T. Pitch	86.7	87.1	87.0	87.2	86.8	85.7	85.1
Model-based	77.6	78.3	78.1	77.1	75.1	72.1	71.4
Binaural	89.4	90.4	84.5	79.0	75.1	70.1	66.9
Proposed	85.1	85.8	86.9	86.4	85.2	82.4	79.7

to the ground truth target or interference pitch points. We generate ground truth target and interference pitch points by feeding the pre-mixed, reverberant target and interference signals into the tandem algorithm. Each detected pitch point from the mixture receives a target/interference label based on its proximity to the ground truth target and interference pitches in that frame. To account for octave errors, if a detected pitch point is not close enough (within a specified threshold) to either the ground truth target or interference pitch points, but a doubling or halving of the detected value is close enough, we label that pitch point according to the doubled or halved value. The label of the contour takes the label of the majority of the frame based labels for that contour.

The model-based system uses pre-trained speaker models to perform sequential organization of simultaneous streams for voiced speech [27]. Speaker models are trained using an auditory feature, gammatone frequency cepstral coefficients, and the system incorporates missing data reconstruction and uncertainty decoding to handle simultaneous streams that do not cover the full frequency range. The system is designed for anechoic speech trained in matched acoustic conditions. To account for both the azimuth-dependent HRTF filtering and reverberation contained in the mixture signals used in our database, some adjustments were made. First, we train speaker models for each of the reverberation conditions that will be seen in testing. For each of the 38 speakers, we select 7 out of 10 utterances for training, generate 10 variations of each of these utterances with randomly selected azimuths for each of the 7 reverberation times. This helps to minimize the mismatch between training and testing conditions, although as mentioned above, the impulse responses used in training are different from those in testing. We found this approach to give better performance than feature compensation methods (e.g. cepstral mean subtraction, and cepstral mean subtraction and variance normalization) for mismatched training and testing conditions.

In [27], a background model is used to allow the system to process speech mixed with multiple speech intrusions or non-speech intrusions. Since we focus on the two-talker case, we found that assuming both speakers are known produces better results than using a generic background model. Incorporating this prior knowledge into the model-based system ensures that we are comparing to a high level of performance potentially achievable by the model-based system.

The binaural system we use for comparison incorporates the azimuth-dependent likelihood functions described in Section 4.2, but labels each T-F unit independently. Although the training paradigm for the likelihood functions is different, this is similar to the system in [24], but more appropriate for reverberant environments. This system does not incorporate the simultaneous streams generated using monaural processing. Whereas our proposed system makes a binary decision about each simultaneous stream as a whole, the binaural system makes a binary decision about each T-F unit independently. For the purpose of comparison, we still measure the percentage of correctly labeled mixture energy *within* the simultaneous streams, even though the exclusively binaural approach is able to generate a binary mask for the entire mixture. This comparison is informative in terms of revealing the amount of reverberation necessary to corrupt binaural cues, and indicates the usefulness of monaural analysis.

In Table 1 we show the sequential organization performance of the proposed system, the model-based system, the binaural system and the two ideal labeling schemes. The ceiling performance measurement achieved by ISO indicates the quality of the simultaneous streams themselves. Any decrease below 100% when using the IBM to generate labels at the simultaneous stream level indicates that the simultaneous streams do not exclusively contain units dominated by the same source. At all levels of reverberation the simultaneous streams have, on average, 88.4% correctly grouped mixture energy. In low to moderate reverberation conditions, the error introduced by our system is, on average, about 2.7%. In the two most reverberant conditions, our system introduces about 7% of the error. The performance improvement over the model-based system is significant, ranging between about 9.5-14% relative improvement depending on the amount of reverberation. This is notable, especially considering that the model-based results incorporate prior knowledge of the two speakers contained in the mixture and training on reverberant mixtures. We can also see that in conditions with T_{60} of 0.2 sec. and above, the proposed system outperforms the exclusively binaural labeling. The transition between 0.1 and 0.2 sec. T_{60} marks the point at which the binaural cues become corrupted enough that monaural analysis improves performance. In anechoic and $T_{60} = 0.1$ sec. conditions, the binaural labeling outperforms even the ideal sequential organization. Note that this is possible because the binaural labeling operates on the individual T-F units whereas the ISO is applied at the simultaneous stream level. This confirms why it is so attractive to use binaural cues at the exclusion of monaural cues in constrained acoustic conditions. At $T_{60} = 0.1$ sec., the average direct-to-reverberant energy ratio is over 45 dB, which although not entirely anechoic, still represents very unreverberant conditions. In the test cases that resemble conditions found in everyday life, where average direct-to-reverberant energy ratio ranges from 5.8 dB at 0.3 sec. T_{60} to -4.1 dB at 0.8 sec. T_{60}, the proposed method provides a relative improvement of 14.6% over the exclusively binaural method.

We provide further analysis of the proposed method and the exclusively binaural method in Figure 5. Here we show performance of each approach as a function of spatial separation

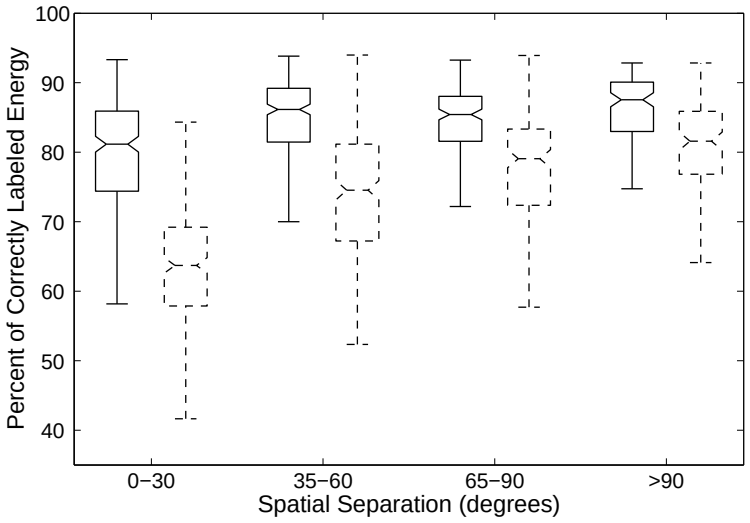


Figure 5: Performance for the proposed method (solid) and the binaural method (dotted) as a function of spatial separation between sources.

between sources. Results are calculated over the 800 mixtures with T_{60} between 0.3 and 0.8 sec. Boxplots are shown for mixtures with spatial separation of less than or equal to 30°, between 30° and 60°, between 60° and 90°, and separation of greater than 90°. The upper and lower edges of each box represent the upper and lower quartile ranges, the middle line shows the median value and the whiskers extend to the most extreme values within 1.5 times the interquartile range. Results for the proposed system are shown with solid lines, and those for the binaural system are shown with dotted lines. We can see that beyond improved performance overall, the proposed method is less sensitive to close spatial proximity. For our system, only the performance of mixtures with 30° or less shows notable degradation relative to the other groups. When using binaural cues alone, the azimuth separation between sources has a much larger impact on performance.

7 Concluding Remarks

The results in the previous section illustrate that integration of monaural and binaural analysis allows for robust localization performance, which enables sequential organization of speech in environments with considerable reverberation. The proposed method outperforms a model-based approach that utilizes only monaural cues and an exclusively binaural approach in all but the least reverberant conditions. We have also shown that incorporation of monaural grouping improves localization performance over two existing methods.

The discrete azimuth space used in this study avoids two potential issues. First, the azimuth-dependent ITD-ILD likelihood functions are manageable in number (37 for each frequency channel in this study). Second, the joint search over all possible azimuths is com-

putationally feasible. In the case of a more finely sampled or continuous azimuth space, or a localization space that includes elevation, one would need to carefully consider how to overcome both issues. To overcome the need for training an unwieldy amount of likelihood functions in a variety of acoustical conditions, parametric likelihood functions could be used without considerable performance sacrifice. In analyzing the trained ITD-ILD likelihood functions, clear patterns emerge that could be utilized to formulate a parametric model. Certain key parameters, such as the primary peak locations and spread of the distributions, could be learned from training data from a discrete set of source positions and extrapolated to a continuous space. The second issue of joint search over all possible angles in a finely sampled or continuous space could be avoided by doing an initial search in a discretized space (such as the one used here), then refining the source positions in a limited range.

The development in Section 5 makes two assumptions that should be carefully examined in future work. First, we propose a maximum likelihood framework in which all sequential organizations are equally likely. For mixtures in which the input SNR is significantly different from 0 dB, maximum a posteriori estimation is more appropriate and it should not be assumed that $P(y)$ is uniform. Second, we assume that all simultaneous streams are conditionally independent. While this may be reasonable for simultaneous streams that are separated in time, this assumption is questionable when two simultaneous streams overlap in time. In the majority of cases, simultaneous streams that overlap in time are due to different sources. Incorporating dependence between simultaneous stream labels should improve performance, but with increased computational cost.

Finally, since the proposed system only processes voiced speech, it is essential to develop methods to handle unvoiced speech. Binaural cues are likely a powerful tool for handling unvoiced speech, which is challenging with only monaural cues (see [19]). Future work must also analyze performance with different types of interfering signals, from multiple speech signals to non-speech intrusions.

Acknowledgment

The authors would like to thank M. Pedersen for providing feedback on a preliminary draft of this manuscript. This research was supported by an AFOSR grant (FA9550-08-1-0155), an NSF grant (IIS-0534707) and a grant from the Oticon Foundation.

References

[1] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, 1979.

[2] J. Blauert, *Spatial Hearing - The Psychophysics of Human Sound Localization*. MIT Press, 1997.

[3] D. Brandstein and M. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[4] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[5] G. J. Brown and K. J. Palomaki, "Reverberation," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. L. Wang and G. J. Brown, Eds. Wiley/IEEE Press, 2006, pp. 209–250.

[6] D. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with an ideal binary time-frequency mask," *J. Acoust. Soc. Am.*, vol. 120, pp. 4007–4018, 2006.

[7] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocky, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 15, no. 7, pp. 1979–1986, 2007.

[8] D. R. Campbell. (2004) The ROOMSIM user guide (v3.3). [Online]. Available: http://media.paisley.ac.uk/~campbell/Roomsim/

[9] H. Christensen, N. Ma, S. N. Wrigley, and J. Barker, "A speech fragment approach to localising multiple speakers in reverberant environments," in *Proc. ICASSP*, 2009.

[10] J. F. Culling and Q. S. Summerfield, "Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay," *J. Acoust. Soc. Am.*, vol. 98, pp. 785–797, 1995.

[11] C. J. Darwin, "Spatial hearing and perceiving sources," in *Auditory Perception of Sound Sources*, W. A. Yost, A. N. Popper, and R. R. Fay, Eds. Springer, 2007, pp. 215–232.

[12] C. J. Darwin and R. W. Hukin, "Auditory objects of attention: The role of interaural time differences," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 25, pp. 617–629, 1999.

[13] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Am.*, vol. 116 (5), pp. 3075–3089, 2004.

[14] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *J. Acoust. Soc. Am.*, vol. 97, pp. 3907–3908, 1995.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. (1993) Darpa timit acoustic phonetic continuous speech corpus. [Online]. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html

[16] W. M. Hartmann, "How we localize sounds," *Physics Today*, pp. 24–29, November 1999.

[17] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," The Ohio State University, Tech. Rep., 2008.

[18] G. Hu, "Monaural speech organization and segregation," Ph.D. dissertation, The Ohio State University, 2006.

[19] G. Hu and D. L. Wang, "Segregation of unvoiced speech from nonspeech interference," *J. Acoust. Soc. Am.*, vol. 124, pp. 1306–1319, 2008.

[20] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 17, pp. 625–638, 2009.

[21] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *J. Acoust. Soc. Am.*, vol. 106, pp. 1633–1654, 1999.

[22] C. Liu, B. C. Wheeler, W. D. O'Brien, R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," *J. Acoust. Soc. Am.*, vol. 108, no. 4, pp. 1888–1905, 2000.

[23] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," MRC Applied Psychology Unit, Cambridge, U.K., Tech. Rep., 1988.

[24] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, no. 4, pp. 2236–2252, 2003.

[25] A. Shamsoddini and P. N. Denbigh, "A sound segregation algorithm for reverberant conditions," *Speech Commun.*, vol. 33, pp. 179–196, 2001.

[26] Y. Shao and D. L. Wang, "Model-based sequential organization in cochannel speech," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 14, pp. 289–298, 2006.

[27] ——, "Sequential organization of speech in computational auditory scene analysis," *Speech Commun.*, vol. 51, pp. 657–667, 2009.

[28] B. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 1986.

[29] S. Vishnubhotla and C. Y. Epsy-Wilson, "An algorithm for speech segregation of co-channel speech," in *Proc. ICASSP*, 2009.

[30] D. L. Wang, "On ideal binary masks as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Boston, MA: Kluwer Academic, 2005, pp. 181–197.

[31] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley/IEEE Press, 2006.

[32] K. W. Wilson and T. Darrell, "Learning a precedence effect-like weighting function for the generalized cross-correlation framework," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 14, pp. 2156–2164, 2006.

[33] J. Woodruff and D. L. Wang, "On the role of localization cues in binaural segregation of reverberant speech," in *Proc. ICASSP*, 2009.

[34] S. N. Wrigley and G. J. Brown, "Binaural speech separation using recurrent timing neural networks for joint F0-localisation estimation," in *Machine Learning for Multimodel Interaction*. Springer Berlin / Heidelberg, 2008, pp. 271–282.